Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-01-05 14:49:46
REFERENCE NO: 171

Author Names & Affiliations

- Harold Garner - Edward Via College of Osteopathic Medicine, Gibbs Cancer Center, Virginia Tech, Orbit Genomics, Heliotext, Comparity, Quanta Lingua, BioAutomation

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

bioinformatics, genomics, genetics

Title of Submission

One stop shop for qualified nextgen “-omics” analysis

Abstract (maximum ~200 words).

Recently, with the advent of nextgen “-omics” data generation techniques that are digital, thorough and inexpensive, there is amassing a tremendous amount of data. For example, it is now cheaper to generate “human genomes” and other “human –omes” than it is to analyze it. Further, the current state of the art requires substantial storage and computation infrastructure, specific expertise and the right choice of analytical approach (from among many).
While there has been an effort to put the data in “the cloud”, there are sever restrictions on accessing this data, bandwidth to move the data to where computations can be done, the proper choice of algorithm/application, output that is interpretable by biomedical scientists, a last generation reference genome (or “-ome”).
What is needed is a dedicated supercomputer center for “-omic” analysis, that has a spectrum of hardware, especially discrete servers with substantial memory, tremendous amount of storage, a host of “qualified” software packages to analyze the data, and a modern reference genome that captures the diversity of the human population. This center will analyze/compare/qualify analysis package performance (computational, but most important statistical and biological rigor) and create pre-configured pipelines with simple interfaces and output so that biomedical scientists can extract real, robust findings from within the data sets they choose to analyze (their own in context with all public data).

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

The challenge is to better analyze “-omics” data in a qualified and uniform way. The size and complexity of the data has led to a tremendous increase in false discoveries and irreproducible results. The goal will be to qualify software and prepare data so that standard analyses are done well, and also enable optimized experimental data analyses. The challenge is to sift through the data and through the many software packages, and provide qualified pipelines for analysis so that complex, data intense analyses can be done by non-programmers, using standard interfaces connected to substantial infrastructure (compute and storage); all as a service.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

The data sets and wide variety of code requirements have made “the cloud” almost useless in this context, for some computations, for example, require servers with modest memory, others with terabytes of RAM, with different economies of parallelization. There are really no technical challenges, for such a center could be built from commodity servers (of different types) and storage. The real challenge is to recognize that analyses is out of control, bordering on unreliable, and certainly is not accessible to most biomedical researchers. We need a layer between biomedical application developers and users, a layer that contains the best (“qualified”) software running on extreme amounts of data by computational unsophisticated users that are blinded from the resources but can count on getting the best results from “qualified” applications such that their output can be compared to others. The goal is to provide a way to unify the computations.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

There are a number of other considerations. All software and pipelines should be analyzed to understand how well they perform, and within what data ranges, settings and nuances. Also, the current human genome has only offers a single base as a reference, completely ignoring that we have two copies of each genome in us, and also ignoring the diversity of states (alleles) available at each base (or amino acid). This diversity can be captured for the human population (thus characterizing the locations that are ethnically diverse) or for an individual (thus characterizing the locations that are mosaic due to accumulated mutations as cells divide). A new reference genome must be built, and could be built to capture this information and make it available to all calculations which have the goal of finding robust markers.

Consent Statement